DOCUMENT RESUME

ED 367 149 FL 021 871

AUTHOR Des Brisay, Margaret; Ready, Doreen
TITLE Defining an Appropriate Role for Language Tests in Intensive English Language Programs.
PUB DATE 91
NOTE 12p.; In: Anivan, Sarinee, Ed. Issues in Language Programme Evaluation in the 1990's. Anthology Series 27; see FL 021 869.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *English (Second Language); English for Academic Purposes; Foreign Countries; Higher Education; *Intensive Language Courses; *Language Tests; *Program Evaluation; Second Language Instruction; *Second Language Programs; Testing; *Test Results; Test Use
IDENTIFIERS Indonesia; Test of English as a Foreign Language; University of Ottawa (Canada)

ABSTRACT

The use of test results as a criterion for evaluating the design of an intensive English language program is examined in one Canadian university program in Indonesia. The study was investigating whether: (1) realistic estimates of training needs were being made; (2) data from other tests would enable better prediction of student outcomes, and therefore of student admission criteria; and (3) some guidelines could be established for balancing test preparation and post-course language use in classroom instruction. Subjects were 129 students in 8 academic English classes. Results from one administration each of the Test of English as a Foreign Language (TOEFL) and the Canadian Test of English for Scholars and Trainees were used to predict gain in scores on a second TOEFL administration. Results show the entry TOEFL scores to be misleading as to the homogeneity of the group and to provide an inadequate baseline for assessing student progress. Implications are found for both the design of the training program and for the use of gain scores in program evaluation. Several Canadian initiatives in English second language training and testing are also described. (MSE)

DEFINING AN APPROPRIATE ROLE FOR LANGUAGE TESTS IN INTENSIVE ENGLISH LANGUAGE PROGRAMS

Margaret Des Brisay
Doreen Ready

In the well-known play, "A Man for All Seasons", which tells the story of Sir Thomas More, there is a scene in which More advises a student of his to become a teacher. " You'd be a good teacher" he says to the young man. The young man replies "And if I were, who would know?"

"Who would know?" More specifically, "How would he know?" Researchers today collect a wide range of qualitative information to answer this question - teacher interviews, classroom observation, peer review, student course critiques - but when interest focuses on the effectiveness of the teaching in terms of measurable educational outcomes, then attention must be given to what the students can do as a consequence of the educational treatment they have received. Pretest-posttest measures of proficiency gains are frequently used to provide quantitative data for evaluating teaching effectiveness, not perhaps in the case of individual teachers but of groups of teachers and the programs with which they are associated. This is particularly true in the informal type of evaluation that goes on when funding agencies or their advisors select the in-country language school that will deliver what is known as pre-departure ESL training to their scholarship candidates. Although teachers and educational researchers are aware of the potential for the misuse of such data, to the lay person, gain scores seems the most the obvious way to determine the success of a teaching program.

In practice, even calculating gain scores seems unnecessarily complicated for many administrators. They prefer to look at pass rates which is understandable given that tests are used primarily to make decisions about individuals. There is some cause for concern, however, when language schools announce in their brochures that 78% of their students "successfully completed their course" which no mention being made of what criteria were used for measuring success and when these were met on schedule or not. It could simply mean that 78% lived to tell the tale...eventually. One formal evaluation of a program with which we were associated featured a chart showing the percentage of students meeting the exit standards in several consecutive semesters. This chart was used to compare the performance of different directors although it made no mention of changes in the length of the semesters, a steady rise in the entry level of the students and compensatory adjustments made to exit scores. If pass rates are to be used to compare programs, they must be interpreted with reference to entry level and other baseline data which, in effect, leads you back to gain scores.

While we concede that test scores have a legitimate role to play in evaluating teaching effectiveness, what follows is , in fact, a cautionary tale, the moral of which is that test scores are not as simple, clear and conclusive as advocates might wish to believe.

CONTEXT FOR STUDY

For the past two years, researchers at the University of Ottawa have been conducting analyses of ESL test scores obtained by students prior to and during their training in several different intensive English programs in Jakarta, Indonesia. The examinees in all cases were candidates for scholarships to either the United States or Canada who had satisfied all the selection criteria for such an assignment except for the English language requirement. This language requirement was defined as achieving a level of English proficiency adequate for academic purposes and was re-stated in terms of a test score of 550 on an International TOEFL.

In other words, not only was the TOEFL being used to measure attainment of the instructional objectives, in many ways, the TOEFL was the instructional objective.

The initial Canadian ESL program in Indonesia had produced disappointing results with many fewer candidates than anticipated reaching the predicted TOEFL score within the allotted time and some being dropped from the program entirely. Moreover, the broadly-defined objectives of the program, preparing students for study abroad, frequently conflicted with the narrowly-defined objective of getting 550 on the TOEFL. However, as one ministry official said when it was suggested that the future needs in an academic environment should be given priority in the program, "Before they can succeed, they must be admitted, and before they can be admitted, they must have 550 on the TOEFL."

The study to be reported in this paper was one of several undertaken in order to try to address the concerns of Canadian program planners about the Indonesian training model. Specifically planners were interested in knowing whether 1) realistic estimates of training requirements were being made 2) whether data from other tests, in particular, the Canadian Test of English for Scholars and Trainees, would enable better predictions about end-of-course success to be made, (and hence, better initial selection) and 3) whether some guidelines could be established for striking a balance between test preparation and preparation for life after the test.

(The comparability study between the two tests has been reported on elsewhere and will be referred to only briefly in the present paper. And it should be stressed that there was never any intention of using the results of this study to evaluate the centres or teachers concerned but rather to use then in setting reasonable objectives for future programs. What sort of gains is it reasonable to expect? Who is most likely to succeed in the time allotted? How many are likely to succeed?)

METHODOLOGY

There were 129 subjects in the study , spread over 8 classes (average of 16 students per class) in three different language centres. Two classes were at Centre A, three at Centre B and three at Centre C. All subjects were studying at the EAPII level. A wide range of test data was collected but the discussion will focus on the test scores from two tests, an International TOEFL written in October 1988 and a second International TOEFL written in January 1989. The October tests were written after students had been studying in a TOEFL preparation program for 12 weeks and January TOEFL following an additional 11-12 weeks of instruction that emphasized academic skills.

Subjects could not be randomly distributed among the three language centres. It appears, however, that students came from similar cultural, linguistic and educational backgrounds. Twenty-six had been studying in intensive ESL programs since January 88, fifty-seven had begun studying at the EAPI level in April and forty-six had tested in when the EAPII program began in July. (No information was available about the previous ESL training of this latter group.) TOEFL entry scores indicated that the range of scores was similar at each of the three language centres although subsequent analysis showed this apparent homogeneity to be a bit misleading. (See discussion below.)

INSTRUCTIONAL TREATMENT

The instruction at each centre was guided by the same general objectives (academic readiness and TOEFL preparation) but , in principle, no single methodology was imposed. An examination of the end-of-course reports does suggest a differential emphasis on TOEFL preparation. One centre (Centre A) administered scores from 22 institutional TOEFL's with each TOEFL followed by a thorough post mortem, implying a minimum of 20% of classtime was devoted to actual TOEFL practice. The fluctuating nature of the scores must have brought tears of frustration to all concerned although they would not have impressed anyone

cognizant of the standard error of measurement and the fact that measurable gains in overall proficiency are not registered weekly. Centre B reported administering 13 practice TOEFL's, mostly in the weeks prior to the October TOEFL while a third (Centre C) reported only three although students were known to have done a lot of practice tests on their own. Two centres give detailed information and assessments from a writing course which took up 25% of classtime. All in all, enough differences were reported in course content that it was decided to further analyze the data to investigate the impact, if any, of these differences on the test scores.

AN EXAMINATION OF REPORTED SCORES

TABLE ONE: Overall Results on TOEFL Total and Part Scores: (n=115)

	JULY	OCT.	JAN.
Total Score	499.0 (18.0)	527.0 (32.6)	533.6 (27.0)
Part One Listening	47.9 (3.1)	52.5 (4.6)	52.6 (2.8)
Part Two Structures	50.3 (3.5)	53.5 (4.7)	53.6 (3.6)
Part Three Vocal/Read	51.3 (3.3)	53.4 (3.6)	54.7 (3.6)

(-) = sd

Table One shows means and standard deviations for totals and part scores on the three TOEFL's for the 115 students for whom there were no missing data. The standard deviations for the group show that the relative homogeneity suggested by the July scores is not maintained in October nor in January. In fact, the standard deviation (an indication of the range of scores) almost doubles. It should be noted that the "July" TOEFL scores were obtained on institutional TOEFL's that had been written at different times.

It can be seen that the mean gain on total scores was 34 points, this being composed of a gain of 28 points during the first 12 weeks which were largely devoted to TOEFL practice at all three centres delivering instruction and 6 points during the remaining 12 weeks where the emphasis shifted to academic preparation. Thirty percent of the students actually achieved the exit standard of 550 on the October TOEFL while 31 percent did so in January.

The standard deviations of 32.6 and 27.0 for the two International TOEFL's in Table One indicate a great deal of individual variation. Individual changes ranged from a gain of 47 points to a loss of 50. In fact, 37 students actually had lower scores in January.

FURTHER ANALYSES

A regression analysis was performed using the Part TOEFL scores obtained in October and the part CanTEST scores obtained at the same time in order to try to arrive at the best equation for predicting the TOEFL scores obtained in January. The analysis excluded the 39 subjects who had obtained 550 on the TOEFL of October since it was felt that they might not be as motivated as the others.

The data were further analyzed using multivariate analysis of variance in order to see if there were any statistical differences among the three centres delivering instruction with regard to the part TOEFL scores on January 22nd. (This procedures also took into account differences in Part TOEFL scores obtained on October 14).

The analysis was repeated using the eight classes as the independent variables instead of Centres to see if classroom variables might have affected gain scores.

REGRESSION ANALYSIS

The results of the regression analysis in which the part CanTEST scores and the part TOEFL scores (October88) including the Test of Written English (TWE) score were used to try to predict the January TOEFL score are reported in Table Two. The method used was stepwise regression in which variables are entered into the equation until there is no further increase in multiple R.

Although this and similar studies (Des Brisay, 1989) show reading scores to account for more of the shared variance than any of the other predictors, their relationship to the dependent variable is not strong enough for them to be used with any confidence to predict success. Moreover, the variables entered into the equation only account for about half the variance present. The total variance present is a measure of how much individual scores vary from the group average. This means that although the prediction equation in **Table Two** gives some idea of what the final TOEFL scores will be there is still a large amount of error so that ESL program planners cannot count on scores obtained before training to give a really accurate prediction of the outcome of training.

TABLE TWO: Regression Analysis to Predict Final TOEFL Score (January) from previous Part TOEFL Scores and Part CanTest Scores(October).

DEPENDENT VARIABLE: TOEFL TOTAL (JAN 89) (N=129)

Predictors	b
TOEFL Reading	3.137
TOEFL Structures	1.806
TOEFL Listening	1.446
Constant	193.887
Multiplier	.717
R^2	.514
Standard Error	16.58

(These results can be contrasted with those obtained when a similar analysis was done in another program (Des Brisay, 1989) where CanTEST was being used in the decision making and incoming students had had limited experience with the TOEFL. In this case, it was CanTEST reading scores which were the best predictors of final TOEFL scores.)

GROUPS FORMED ON DATE OF ENTRY AND ENTRY SCORES

In order to see whether it might be possible to control some of the sources of individual variation, the data were examined to see if differences in either gain scores and/or pass rates could be related to individual differences in proficiency at entry or the length of intensive training as measured by date of entry into the program. Descriptive statistics for groups formed by date of entry and by TOEFL1 (July) scores are shown in **Tables Three and Four.**

In **Table Three**, we see that students who tested directly into EAPII(Group 3) had higher means and more successes than those who were promoted in from EAPI (Group 2) while these in turn had better test performances than those who had previously done both BELT and EAPI (Group 1). (Only the differences between this latter group and the direct entrants were significant and then only for the listening and reading sections.)

Table Three: Means and Standard Deviations on Part TOEFL Scores for Groups Formed on Date of Entry.

	October	January	Gain
Group One: 01/88 (n=29)	515.64 (33.8)	527.05 (27.6)	11.4
Group Two: 04/88 (n=57)	521.75 (32.2)	530.9 (27.9)	9.3
Group Three: 07/88 (n=52)	539.8 (28.7)	540.51 (26.7)	.71

Table Four: Means and Standard Deviations on Part TOEFL Scores for Groups Formed on July TOEFL Scores.

	October	January
Group One:(below 500) (n=69)	520.80 (29.36)	530.05 (28.11)
Group Two: (501-525) (n=51)	541.6 (27.4)	542.79 (24.52)
Group Three: (over 525) (n=14)	549.33 (23.4)	548.66 (23.05)

However, when we look at the gain scores by date of entry we see that the students who began their intensive instruction in January are making larger gains even though they are still farther below the exit standard; they are making larger gains as a group partly just because they are weaker and students in the lower score ranges typically register larger gains.This difference by level reflects the fact that test scores are not truly equal interval in terms of knowledge increment. An comparison of the gains made by three groups formed on the basis of their initial TOEFL scores (Table Four) supports this in that larger gains are observed among the lower proficiency groups.

There would be no way to further explore sources of individual differences without more knowledge of the previous language learning experiences, general intelligence and particular learning styles of this group of students. (Scores from an academic proficiency test were available and correlated at .07 with TOEFL entry scores and .33 with January TOEFL scores).

CENTRES AND CLASSES AS VARIABLES

Table Five: Means and Standard Deviations for Part TOEFL Scores October and January by Centre.

CENTRE	Listening Oct	Listening Jan	Structures Oct	Structures Jan	Reading Oct	Reading Jan
A (n=31)	52.4 (4.1)	53.2 (3.5)	54.3 (4.4)	54.7 (3.5)	54.1 (3.8)	55.8 (3.6)
B (n=50)	51.2 (5.0)	51.8 (4.4)	52.4 (5.5)	52.9 (3.3)	51.3 (3.7)	53.3 (3.5)
C (n=48)	53.2 (4.4)	52.6 (3.0)	52.3 (4.4)	53.1 (3.8)	54.0 (3.2)	54.4 3.5)

The results of the multivariate analysis do not support any conclusions about the efficacy of any one Centre over another. The analysis does support the conclusion that Centre A is stronger but this is true in October as well as January. There were no significant gains made by any Centre on the listening and structure section of the test and all three made gains in reading which did not differ significantly from each other.

The matter of statistical significance is very important considering the decisions that may be made on the strength of the appearance of differences. It should be kept in mind that group average scores, such as those in Table Five, are made up of individuals scores which may vary considerably from the average score. These individual scores , depending on whether they are well above or below that average score, can raise or lower it accordingly. Thus, although there may appear to be between group differences , once the group mean scores have been analyzed using rigourous statistical methods, these differences become something attributable to chance alone; in other words, the differences are not statistically significant.

This lack of statistical significance is not entirely unexpected especially in view of the fact that students are never randomly distributed to training groups and there is no guarantee that the groups being compared were ever equal before training began.

Table Six: Means and Standard Deviations for Total TOEFL Scores October and January by Class

CLASS	JULY	OCT.	JAN.	Raw Gain	Pass (%)
1 (16)	491.0 (11.6)	532.0 (27.6)	543 (23.7)	11	50
2 (17)	489.0 (12.9)	540.0 (29.8)	546.0 (32.4)	6	44
3 (18)	503.0 (19.3)	507.0 (36.3)	523.0 (30.5)	16	17
4 (16)	495.0 (20.6)	509.0 (31.8)	519.0 (23.7)	10	12
5 (14)	511.0 (17.2)	537.0 (34.9)	538.0 (26.8)	1	50
6 (16)	496.0 (17.3)	519.0 (36.4)	533.0 (26.3)	14	44
7 (17)	504.0 (19.8)	542.0 (22.9)	538.0 (23.7)	-4	13
8 (18)	501.0 (17.8)	530.0 (26.2)	529.0 (28.4)	-1	23

Table Six gives similar statistics for all eight classes. As previously noted, the scores for TOEFL1 were obtained on institutional TOEFL's and were not all written at the same time. However, all students wrote a version of the Canadian Test of English for Scholars and Trainees within the first week of their program and the classes are ranked in a similar way according to the CanTEST results. As was the case with Centres, none of these differences is statistically significant.

Some researchers would question whether raw gains should be used at all to measure growth in instructional settings, much less to make comparisons among different groups since raw gains typically level off as students become more proficient. Swinton (1983) describes one possible source of error in calculating gain scores when there is a wide range in scores. That is the statistical phenomenon of the regression to the mean. With this data, that is not a threat because the range of scores is extremely narrow (475-525)

IMPLICATIONS OF THE REPORTED SCORES FOR THE TRAINING MODEL

As stated initially in this paper, the data were not collected for the purpose of evaluating the teaching at the different centres but rather for evaluating the training model itself. In this context, the study clearly shows the need for better baseline data. The means and standard deviations for the entry TOEFL (Table One) give a misleading impression of the homogeneity of the group, even allowing for the imperfect way this can be reflected in any test score. Although it is commonly found that students will progress at different rates so that the range of abilities in a given group may increase over a period of instructions, nevertheless, the July scores, which were obtained on a number of different institutional TOEFL's written at different times, do not provide adequate baseline data for determining progress. This is a finding that can be easily operationalized. However, a more controlled selection process should not be undertaken for the purpose of keeping people out but for making more realistic estimates of training requirements.

Improving the pass rate within the present time frames would involve insisting on higher entry scores. This too could be easily operationalized but would seriously reduce the pool of potential candidates and risk putting concerns about costs of language training ahead of the larger aims of such technical assistance programs which imply giving an equal opportunity to all otherwise qualified candidates. Moreover, although the perception of the teachers that continuing students are somehow weaker is supported by the findings , this can in no way be interpreted to mean that as a group they are poorer language learners. Their poorer performance simply reflects the fact that as a group, they were only minimally proficient for EAPII on entry and had further to go to reach the exit standard.

The test data do not permit any useful comparisons to be made among the centres involved. The observed score patterns might well be interpreted differently if less refined statistics, such as pass rates or gains on total scores (enlarged by the ETS practice of multiplying everything by ten thirds) were used. In that case, some classes and some centres could appear to have been more successful than others. The percentage of students achieving the desired TOEFL score did vary from class to class (50% to 14%) and centre to centre (45% to 27%) but as we have noted above, following multivariate analysis, none of the gains on part scores shown in **Tables Five and Six** were found to be statistically different by class or centre. The difference in pass rates, then, could be equally well attributed to chance and/or to the characteristics of the class on entry. The extent to which administrators would be impressed or distressed by the score patterns revealed in this study would partly depend partly on their degree of statistical sophistication.

IMPLICATIONS FOR THE USE OF GAIN SCORES IN PROGRAM EVALUATION

In the particular program under study, efforts had been made to avoid the methodological weaknesses that have plagued other attempts to quantify teaching effectiveness. Classes were of similar size with a similar balance of continuing and newly placed students. Instruction was of the same length and intensity, and as previously mentioned, students were of similar educational, cultural and linguistic backgrounds and students were thought to be at similar levels of proficiency on entry.

It is individual differences in proficiency gains as measured by the TOEFL which are the dominant finding of this study. Whatever group tendencies can be found are of limited use in program planning and of virtually no use in program evaluation. We can estimate from this and other similar studies that approximately 1/3 of a group of students studying at the EAP II level will reach TOEFL 550 after 18 to 20 weeks of intensive ESL instruction but which ones they will be cannot be predicted from the test scores. (Probably the teachers know, but how do they know?)

The fact that no statistically significant differences among centres or classes were

found may simply suggest that all centres were equally effective (or perhaps, from a sponsor's point of view, equally ineffective). However, the fact that group means disguise so much individual variation and the testing instrument used failed to measure the learning that must have been taking place does have implications for future efforts to use gain scores as a measure of program effectiveness. Such efforts will have to recognize, as educators have always done, that:

> no treatment can be equally appropriate for everyone and as a corollary to this, similar instructional treatments will have a wide range of outcomes. We may be able to say that 35 to 40% of students entering an intensive ESL program at the EAPII level will reach the exit standard of 550 on the TOEFL with 22 weeks of instruction, but which ones they will be, we cannot say;
>
> general proficiency tests are not appropriate for measuring gains over short periods of time (if 360 hours of intensive training can be considered short) and, moreover, such tests will be particularly insensitive to growth in specific skill areas such as writing for academic purposes;

While educational researchers consistently stress that evaluation cannot be based solely on testing student product, (Weir, 1989), program accountability does seem to require that an instructional program have measurable educational outcomes . Given that it would not be cost-effective to provide individualized instruction and assessment, then clearly more appropriate testing instruments an statistical techniques are required.

Bachman (1986) optimistically declares "new developments in criterion-referenced test theory and more comprehensive definitions of language proficiency provide keys to developing language tests that are appropriate to the needs of language program evaluation." Developing such a testing system takes time and a good deal of money. Even when the reliability and validity of the new instrument has been empirically established, one must still establish its credibility in the eyes of the gate-keepers to North American universities. It becomes a question not of "Who would know?" but "Who would believe you?"

The poor performance of these 129 subjects on the January TOEFL offers a compelling argument against the use of a norm-referenced standardized general proficiency test to measure achievement in an academic skills program. You will recall that there was group gain of only 6.2 points and perhaps the most striking finding in the study is the large number (48 out 129) of students who actually had lower TOEFL scores in January than they had had in October, something that cannot satisfactorily be explained away by referring to standard error and regression to the mean.

It is not unreasonable to assume that the 39 students who had achieved their exit score in October were less motivated to do well on the January TOEFL. Here, as elsewhere, there was great individual variation. Twenty-one of the students scoring 550 or more in October had lower scores on the January TOEFL, 3 remained the same, 12 improved and three others did not (wisely, perhaps) write the second TOEFL. On the other hand, 24 students who had not passed in October also had lower scores in January, a phenomenon not likely to be explained by a decrease in motivation.

Twenty two of these "losers" were students in Program C, the least TOEFL intensive of the three centres. Neither the possibility that the January test was easier or that nothing was learned can be seriously entertained. The fact that the "losers" tend to be concentrated in the centre providing the least TOEFL practice between the two tests and the "winners" in the program providing the most, suggests an attractive line of inquiry that would be impossible to pursue on the basis of the data available. It is tempting to suggest group differences might have been more marked had not the need for achieving a certain TOEFL score been uppermost in the students' minds . Given this pressure to pass the TOEFL, they may have simply selected from the different programs whatever they thought would be useful to them in achieving this end and did not fully engage in the rest.

Even in studies where differences in gains and successes can be shown to be statistically significant, it is difficult to trace causal relationships. To quote Long (1983), "We often don't know if he gained because of the program, in spite of the program or merely while registered in the program." When it comes to choosing an institution to deliver ESL

instruction any informed program planner knows that other information must be collected. Canadian planners, for example, observe classes, examine curriculum documents, evaluate facilities, such as a libraries, provisions for self-study, support staff, language labs, look for resources that will provide cultural and academic preparation in addition to language training and do not allow themselves to be unduly impressed by claims of a high pass rate or promises of dramatic gains. It is not unknown, however, for groups of students to be moved from one centre to another or for individual students to be dropped from a program because improper inferences have been drawn from test results.

CANADIAN INITIATIVES IN ESL TRAINING AND TESTING

The Canadian International Development Agency funds several Human Resources Development Projects that have a language training component. The goal of the latter is to select, train and certify candidates from the developing countries who wish to come to Canada for either university study or practical attachments. As with many similar development programs, planners have had to face the fact that the greater the number of stakeholders, the greater the need for some form of standardized evaluation to provide for comparability among programs and overall program accountability.

Fortunately, they have also come to appreciate the extent to which a certification test can "steer the curriculum" (Canale 1988). In order to ensure positive washback from the test used, CIDA has provided financial support for the development of a program-specific testing system. This is clearly not a solution for everyone. I mentioned some of the problems above. However, CIDA is acutely aware of how inaccurate assessments of ESL proficiency can result in unexpected expenses to the funding agency as well as lost opportunities for otherwise qualified scholars and trainees.

The Canadian Test of English for Scholars and Trainees is compiled from an item bank consisting of authentic texts for both listening and reading comprehension and is supplemented by a writing exam and an oral interview. The fact that more information about language proficiency is available when making the initial selection and that students must be at least at a level corresponding to EAPI means that nearly all non-academic track candidates (trainees) are able to reach their exit standards in an 18 week semester while academic track candidates (scholars) generally require two semesters.

Test reaction questionnaires are completed by both teachers and test takers following each administration. This input, plus continuing dialogue with teaching staff and materials developers help strengthen the alignment between curriculum and tests so the tests can more credibly measure attainment of the instructional objectives.(Gatbonton, 1989, Des Brisay 1989) For example, there are no single sentence prompts, no isolated grammar or vocabulary questions on the CanTEST as teachers found this discouraged students from dealing with longer contextualized samples of language. Finally, the fact that the tests are normed on specific sub-sets of the international test clientele permits a more sensitive interpretation of scores.

We would like to mention briefly three other programs which provide alternative models designed to lessen overdependence on test scores for making consequential decisions. In one program , students are relieved of the necessity of writing any ESL tests beyond the first one. A thorough diagnosis is made at entry and generous training estimates are made. (After all, you can always send someone abroad earlier than anticipated but it is demoralizing to keep him or her back.) No formal testing is done again after the initial projections so that the instruction can focus on preparing students for the future. Although the CanTEST is administered to provide for program accountability (and comparability), decisions affecting the students are not based on CanTEST scores, more or less eliminating test anxiety. Such a program is only possible because a small group (15 per year) of students is involved, special arrangements have been made with the admissions office at their Canadian university and administrators are prepared to offer any necessary post-admission ESL support.

Another program recognizes the fact that language proficiency takes a long time to develop and it may not be cost-effective to keep a student in language training until he has reached a proficiency level adequate for the writing of a Ph.D thesis. In this program students begin their course work in their own country with visiting Canadian professors before coming to Canada for 12 months of study. They then return home to write their theses in their mother tongue. The degrees are joint degrees (Ph.D in management) granted by the Canadian and Chinese universities involved.

Yet another program which allows for the steady but slow maturation of ESL proficiency without excessively delaying academic training involves a teacher training program for future ESL teachers in Malaysian secondary schools. By selecting recent high school graduates for this program, the high cost of removing an active professional from the work force for lengthy periods of language training is avoided. The students do all their undergraduate training in Canada but Canadian faculty are counselled on how to evaluate their work in spite of ESL problems and marks assigned in the first two years make allowances for communication problems related to ESL proficiency.

And finally, recognizing that the information requirements of sponsors and admissions officers will dictate the continued use of standardized tests for certification in most programs, TESL Canada is trying to encourage the informed use of a wider range of ESL admissions tests in Canadian post-secondary institutions through the production of a manual for test score users. The proposed user's guide will explain how different tests relate to each other and contain details concerning the reliability, accessibility, quality, significance and security of the information of each test. The TOEFL, the CanTEST, the new IELTS, the Ontario Test of English as a Second Language (OTESL), the University of Toronto's Certificate of Proficiency in English (COPE) are among the tests to be included. It is hoped that this manual will enable programs which do not have the resources to develop their own test to at least pick the one closest to their needs with confidence that the scores will be recognized by receiving institutions.

In closing let me finish the story of Sir Thomas More and his student. When the student complained that no one would know if he were a good teacher or not, More replied. " You will know, and your students, and God. That's not a bad audience." Unfortunately, the audience does not seem to have enlarged much since More's time and since God is not available to work on evaluation teams, we must look to other authorities to satisfy the information requirements of external stakeholders. This demands new models for evaluating instructional programs in which the role played by test scores must continue to be interpreted carefully and cautiously.

REFERENCES

Bachman, L.B. (1989); The Role of Criterion-referenced Test in Language Program Evaluation,in Johnson,K., eds. .....Cambridge University Press

Canale, M. (1987). Language assessment: The method is the message. In D. Tannen and J.E. Alatis (eds), The interdependence of theory, data, and application, Washington, DC: Georgetown University Press. 249-262. (Georgetown University Round Table).

Canale, M. (1988). The Measurement of Communicative Competence, in Annual Review of Applied Linguistics, 8, 67-84. Cambridge University Press.

Des Brisay, M. (1989). The Problem of the Middle Ground: where do you draw the line? Paper presented at the 11th Annual Language Testing Research Colloquium, San Antonio, March 1989.

Gatbonton, Elizabeth (1989). Taxonomy of Language Tasks and Objectives: English Course for Chinese Visiting Scholars, Trainees and Consultants in Canada. Document prepared for the Canada China Language and Cultural Program, St. Mary's University, Halifax.

Long, Michael,(1983). Presentation at TESOL Summer Institute , University of Toronto , Toronto, July 1983.

Overseas Training Office, Government of Indonesia(1988). Preparing Indonesians for Overseas Training. Presentation at the National Association for Foreign Students Affairs Conference, Washington, DC. June 1988.

Swinton,D. (1983). Measuring Growth in Instructional Settings.TOEFL Research Report No.13.Educational Testing Service, Princeton,

Weir, C. (1989); Program Accountability;The writing is on the wall; unpublished document, Reading University .